Mini-Review

Lot-to-Lot Variation

Simon Thompson,[1] *Douglas Chesher[1,2]

[1]Chemical Pathology, NSW Health Pathology; [2]Northern Clinical School, University of Sydney, Royal North Shore Hospital, St Leonards, NSW 2065, Australia.

*For correspondence: Dr Douglas Chesher, doug.chesher@health.nsw.gov.au

Abstract

Lot-to-lot variation affecting calibrators and reagents is a frequent challenge that limits the laboratory's ability to produce consistent results over time. This variation is not without clinical consequence and there are several well-documented examples of adverse clinical outcomes. It is important that laboratories have procedures in place for quantification of this inaccuracy, and for determining whether the amount of variation is acceptable for the release of patient results. Various approaches have been taken to the assessment of new lots, including the evaluation protocol published by the Clinical and Laboratory Standards Institute (CLSI). Internal quality control and external quality assurance material is often not commutable, and so the use of native patient samples is preferred. Published evaluation protocols differ significantly in ease of use and statistical rigour, and some may be underpowered to detect a clinically meaningful change between lots. Furthermore, current protocols (including the CLSI protocol) will not detect cumulative shifts between reagent lots. This shortcoming may at least partly be addressed by laboratories adopting moving patient averages or similar quality procedures. Collaboration and data-sharing between laboratories and manufacturers also has an important role to play in the detection of lot-to-lot variation. While the laboratory may take steps to evaluate and detect variation, the ideal is to reduce variation between lots at the point of manufacture. Using appropriate acceptance criteria based on medical need or biological variation requirements instead of some arbitrary percentage may go some steps toward achieving this.

Introduction

It is essential that laboratory results are consistent over time to enable clinicians to interpret these against reference intervals and past values. Many factors may affect the reproducibility of results in individual patients but one that has challenged laboratories for many years is variation between different manufacturing lots of calibrators and reagents leading to shifts in patient results and quality controls.[1] Lot-to-lot variation (LTLV) has been increasingly recognised as an important source of analytical error, and is a factor that should be considered when determining whether an assay is fit for purpose. In this article, we will first describe the problem that is LTLV, before looking at current approaches which allow us to quantify LTLV and determine whether a new reagent lot is acceptable for use. We will finish with a discussion of the limitations of current evaluation protocols, and suggest how these limitations could be overcome in the future.

Why Does Lot-to-Lot Variation Occur?

In an ideal world, each lot of reagent and calibrator produced by a manufacturer would be identical, which would allow for the laboratory to seamlessly transition from one lot to the next with no noticeable change in patient results. Unfortunately the realities of the reagent preparation process mean that there will always be some differences between reagent lots. These differences tend to be more marked in immunoassays than for general chemistry assays. Production of an immunoassay reagent involves the binding of antibodies to a solid phase. The quantity of antibody bound to the solid phase will inevitably be slightly different for each batch of reagent, even when external factors such as temperature, pH and concentrations of the reagent constituents are kept consistent.[2]

Manufacturers do have internal quality control procedures which aim to detect variation between lots of reagent,[1] however it may be that these procedures are inadequate at detecting clinically significant shifts in reagent performance.[3] The criteria used by manufacturers when determining whether a lot is appropriate for use are frequently arbitrary and do not reflect the updated Milan criteria for defining performance specifications;[4] therefore a laboratory receiving a new lot of reagent should not rely on manufacturer evaluation data.

Clinical Consequences of Lot-to-Lot Variation

Clinically significant LTLV, when undetected, can cause changes in results which may present a risk to patient care. For example, Thaler *et al.* report a case in which a change in HbA_{1c} reagent lot led to an average increase in patient results of 0.5%.[3] A change of this magnitude may lead to patients being incorrectly diagnosed with diabetes mellitus and erroneously started on medication. Undetected LTLV has also been reported for insulin-like growth factor 1 (IGF-1).[5,6] In the case described by Algeciras-Schimnich,[5] the discrepancy went unnoticed despite the laboratory utilising an evaluation procedure for new reagent lots, and was brought to the attention of laboratory staff by clinicians noticing an unusually large number of discrepant results. LTLV affecting several lots of prostate-specific antigen (PSA) reagent in use at SA Pathology was responsible for the release of falsely-elevated PSA results, causing undue concern for patients who had undergone prostatectomy (as such a result would be suggestive of a recurrence of prostate cancer).[7]

LTLV is not a problem restricted to the biochemistry laboratory. LTLV has also been described in other branches of laboratory medicine including haematology and serology.[8,9] Procedures for the identification of LTLV will vary depending upon the nature of the analyte in question, and therefore processes developed for other disciplines may not be appropriate for use in a biochemistry setting.

When Should LTLV Evaluation be Carried Out?

With every change in lot of reagent or calibrator, there is the potential that clinically significant LTLV may be present; therefore a full evaluation of each new lot should ideally be carried out. Furthermore, it is a requirement under ISO 15189 that each new lot or shipment is acceptance-tested prior to use.[10] However, it must be noted that some laboratories struggle to fulfil this requirement. The necessity for 'just in time' reagent ordering practices, as takes place in some of the larger laboratories, means that it can be difficult to complete a thorough evaluation of a new lot prior to it being put into service.

Evaluation is usually not required when changing to a new bottle of reagent or calibrator from the same lot as the constituents of each bottle within a lot should be almost identical, with a negligible impact on patient results. If there is significant vial-to-vial instability this can be checked with internal quality control.

Limitations of LTLV Detection Using IQC or EQA Material

Evaluation of a new reagent or calibrator lot involves the measurement of a sample on both lots, followed by statistical analysis of the paired results to determine whether the new lot meets pre-determined acceptability criteria. It may be difficult to obtain sufficient volumes of patient sample across the measuring range of the assay, and so internal quality control (IQC) and/or external quality assurance (EQA) material is sometimes used as a substitute for patient samples. However, there is considerable evidence in the literature demonstrating the poor commutability of IQC and EQA material when evaluating new reagent lots.[11-14] In one large study spanning several instrument platforms and analytes, there was a significant difference between the results obtained for IQC material and patient serum in 40.9% of reagent lot change events.[15]

The poor commutability between IQC/EQA material and patient samples means that a change seen with IQC/EQA material may not be present when the matrix is patient serum. This could lead to inappropriate rejection of the new lot. Of more concern is the possibility that a significant change in patient results would not be identified using only IQC/EQA material for comparison, which may result in inappropriate acceptance of a new lot, and the potential for inaccurate patient results. It is therefore recommended that fresh patient serum be used when evaluating new lots of reagent. Commutability issues are relevant when a new lot of calibrator is being assessed with the same reagent.

Having to perform patient comparison studies with each new lot creates a significant burden both in time and cost for the typical clinical biochemistry laboratory. In order to reduce this workload, Martindale *et al.* have recommended categorising assays into three groups and using these categories to decide whether assessment must be made with patient samples or whether the use of IQC material will be sufficient.[16] The first group are those assays where evaluation by measurement of IQC is the only practical method due to analyte instability, or where a test procedure is particularly labourious, such as faecal fats. The second group are those assays where analysis of historical data shows that changes in patients results and IQC are clinically unimportant, typically associated with shifts in IQC of less than one standard deviation (SD). The third group are those that have demonstrated clinically significant shifts in patients or controls. With the first and second groups, they recommend initial assessment by IQC alone but for the third group of assays, acceptance testing using patient samples is required. Sharing of data within a laboratory network using the same methods also has the potential to reduce the burden on individual laboratories.

The Evaluation Process

Whilst specific protocols for evaluation of LTLV vary between laboratories, the general principle remains the same, and is

1. Acceptance criteria are determined in advance.
2. An appropriate number of samples are selected for evaluation.
3. Samples are tested on the old lot and the new lot.
4. Results between the lots are compared.
5. If the results satisfy the acceptance criteria, the new lot is accepted. If the results do not satisfy the acceptance criteria, the new lot is rejected and must not be used for the measurement of patient samples.

Actions to take if alternative lots are not available include the issuing of comments to clinicians with result reports, the amendment of reference intervals and the implementation of factors so that the new results are comparable with previous results.

Figure 1. The stages of a LTLV evaluation process.

outlined in **Figure 1**. The first stage of the evaluation is to determine the criteria which will be used to decide whether the new lot is acceptable or not. These criteria can be determined in a number of ways, as discussed in the next section of this review. Once the acceptance criteria have been defined, it is necessary to determine the number of patient samples to be used for the comparison. Increasing the number of samples will increase the statistical power of the evaluation, meaning that there will be a higher probability of successfully detecting a clinically significant shift in results. It may be that clinically significant LTLV is only present for certain concentrations of analyte, and therefore it is recommended that the study samples span the analytical range of the assay where possible.

Testing of each sample should be carried out on the same day, using the same instrument and the same operator. Once testing is complete, statistical analysis of the paired results will allow the evaluator to compare with the acceptance criteria and make a decision as to whether the new lot is acceptable for use.

As outlined above, commutability issues with IQC and EQA materials frequently mean that a LTLV evaluation cannot be adequately carried out using these materials alone. However when switching to a new lot of reagent or consumable, it is also important to investigate the performance of IQC material with the new lot. This is normally achieved as part of a precision study, in which it is generally appropriate to substitute IQC material for patient samples. In some cases, there may be no evidence of clinically significant LTLV using patient material, although the new lot does provide different results for the running mean and/or SD with IQC material. In these situations one can safely alter the IQC targets for the new reagent lot. However, a large difference in IQC performance in the absence of a noticeable difference with patient material should prompt scrutiny of the LTLV evaluation procedure – it may be that the current procedure is statistically underpowered and is failing to pick up a clinically important shift in patient results.

Determining Appropriate Acceptance Criteria

Prior to measurement, it is important to determine the criteria which will be used to decide whether the new lot has passed or failed the evaluation. These criteria may take the form of a 'critical difference' which is considered to be the maximum possible difference between results which would not adversely affect clinical outcomes. Other acceptance criteria may also be used, depending upon the evaluation protocol. Examples of acceptance criteria can be found in **Figure 2**. Evaluation protocols may specify that more than one criterion be met before the new lot is accepted for use.

Determining acceptance criteria appropriate for each analyte can be challenging. The CLSI recommend choosing criteria based on the Stockholm hierarchy, published by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) in 1999.[17] This has been superseded by the updated Milan criteria for defining analytical performance specifications,[4] and so it seems appropriate to use these criteria as a guide when determining acceptable differences between lots. The Milan criteria propose three models:

(i) criteria based on clinical outcomes
(ii) criteria based on biological variation data
(iii) criteria based on state-of-the-art performance.

Whilst it is best to select acceptance criteria based on clinical outcomes studies, this data is often unavailable. Therefore it may be impossible to achieve this top level of the hierarchy.

- A 'critical difference' between the two lots.
- Slope and intercept values for a regression equation between the two lots.
- R^2, based on Pearson's correlation coefficient.
- Mean difference between the lots of reagent/calibrator (paired t-test).
- Individual differences between paired measurements.
- Allowable Bias (Ba), based on biological variation data.
- Total allowable error (TEa), based on biological variation data.

Figure 2. Examples of acceptance criteria used in LTLV evaluation protocols.

Biological variation data exists for many commonly measured analytes and is easily accessible online. This data can be used to calculate the total allowable error (TEa), and the acceptance criteria may be set at a certain percentage of the TEa.

In circumstances where biological variation data is unavailable or inappropriate for use, acceptance criteria can be chosen based on current state-of-the-art performance. Sources for this data include the analytical performance specifications of EQA programmes or data from a direct comparison with other laboratories offering the same test.

Whilst the Milan hierarchy provides a good framework for deciding upon a critical difference value for the comparison, other factors such as the values obtained by regression analysis between the two lots may be used as additional criteria. Additional acceptance criteria may increase the likelihood that the comparison identifies a clinically significant difference between the old and new lots.

Selection of Patient Samples for Evaluation

After appropriate acceptance criteria have been decided, it is necessary to determine how many patient samples will be included in the evaluation exercise. Target concentrations should be focussed around medically-relevant decision limits for the analyte, and should span the assay range. A minimum of two target concentrations should be used, although more target concentrations will result in a more thorough evaluation.

The number of samples at each target concentration to be tested is influenced by a number of factors including the imprecision of the assay, the magnitude of the desired change to be detected and practical considerations such as the ability to obtain an appropriate number of samples. This is discussed in more detail in the next section of this review. The CLSI guidance on evaluation of LTLV recommends that, as a minimum, a total of three separate samples are used across the assay range.[18] Analysis of patient samples may be done in singlicate or replicates depending on the protocol being used, as will be described below. Increasing either the number of samples or the number of replicates will improve the statistical power of the evaluation.

It is important to ensure that there is enough volume of each sample in order to complete the evaluation. This should include extra volume to accommodate any samples which need to be rerun. Samples should be handled in accordance with the requirements of the reagent manufacturer, and should be free from interfering substances such as haemoglobin, bilirubin and excessive endogenous lipids.

It may be difficult to obtain samples for target concentrations which rarely occur in practice yet remain clinically important. One way to deal with this problem is to actively collect and freeze spare patient samples which meet these criteria when they come into the laboratory, so that they are available for use when the time comes to carry out an evaluation. Alternatively, laboratories may share such samples with other laboratories within the same network.

If it is not possible to obtain enough volume from one sample, pooled samples of a similar concentration may be created. Once a pool has been created, it can be split into separately-frozen aliquots of adequate volume. This allows for only one aliquot to be thawed per evaluation, and will prevent adverse effects due to repeated freeze-thaw cycles on the pooled sample.

Patient samples should be stored in a frozen state. CLSI guidelines recommend that samples are stored at or below -70 °C which will minimise their degradation during storage.[18]

CLSI Evaluation Protocol 26-A

In 2013, the CLSI released guidance on conducting LTLV evaluation studies.[18] The advice on selection of samples and acceptance criteria outlined above are influenced heavily by this guidance. A suggested evaluation process is also included in the CLSI guidelines. This process is thorough and would be expected to detect the majority of clinically significant shifts in reagent lots; however it does represent a significant work burden on the laboratory, and may be unfeasible for some laboratories.

The evaluation process recommended by CLSI EP26-A is outlined in **Figure 3**. This protocol uses the principle of a 'rejection limit' (RL), which is defined as a specific percentage of the pre-determined critical difference. It is also necessary for users of this protocol to decide upon the desired statistical power required for the comparison. The statistical power is the probability that a statistically significant difference between the lots will be detected by the evaluation. For most purposes, a statistical power in the range 0.8–0.95 would be appropriate (a statistical power of 0.8 indicates that there is an 80% chance of a statistically significant difference being detected).

The statistical power chosen will vary depending upon the analyte in question, and users must give consideration to the likelihood of a small change in results impacting upon medical decision-making. For analytes where a small change has the potential to adversely impact clinical decisions, it is advisable to opt for a higher statistical power.

CLSI EP26-A also takes into account the analytical imprecision of the assay. This is represented as both the repeatability (S_r) (also known as within-run imprecision), and the within-reagent-lot imprecision (S_{WRL}). In most cases, S_{WRL} can be substituted for the between-run imprecision obtained from IQC data for the assay. A precision study is normally conducted as part of the evaluation of a new lot of reagent, and S_r and S_{WRL} can be obtained from this study. If the laboratory does not have access to their own precision data, an alternative source of this information may be the assay manufacturer's instructions for use.

The values for S_r and S_{WRL} may differ across the assay range, so users of the CLSI protocol must ensure that they use data relevant to each target concentration of patient sample measured.

Once the above data is obtained, using lookup tables in the CLSI document it is possible to ascertain information about the number of samples required at each target concentration, the actual statistical power obtained, and an appropriate rejection limit. Samples are collected and analysed on each lot of reagent. The mean difference at each target concentration is compared against the RL. If all differences are less than the RL then the new lot is deemed acceptable and can be used to provide patient results.

- Precision data is used to determine the repeatability (S_r) and within-reagent-lot standard deviation (S_{WRL}).
- Desired statistical power is determined.
- Lookup tables are used to determine:
  - » the number of samples required at each target concentration
  - » the rejection limit (which will be a percentage of the critical difference)
  - » the actual statistical power obtained by the evaluation
- Samples are selected and measured on each reagent lot.
- The mean difference between results for the current and new lot at each target concentration are compared with the rejection limit:
  - » If all differences are less than the rejection limit then the new lot is accepted for use.
  - » If any of the differences are greater than the rejection limit then the new lot has failed the evaluation, and must not be used until further troubleshooting has been carried out.

Figure 3. CLSI procedure for the evaluation of new reagent lots.

The CLSI protocol is statistically robust and, when used correctly, it would be expected to identify the vast majority of clinically significant shifts between reagent lots. However it is not without its problems. Firstly, it relies upon the appropriate selection of a number of variables, including a critical difference and statistical power. If these values are not appropriate then the ability of the protocol to identify LTLV is greatly compromised. Secondly, it requires a significant amount of work to carry out in full. A busy laboratory with many different analytes and lot changes each year may not have the time or resources to follow the protocol thoroughly. Thirdly, the lookup tables in the document assume that an alpha error value of 0.05 is chosen. This is normally adequate unless a higher degree of certainty is desired, in which case it may be desirable to opt for a lower alpha level. In such cases the lookup tables are invalid. Finally, this protocol may recommend that an unfeasibly large number of patient samples are used for the study. In these situations, increasing the number of replicates of each sample may be an alternative way to obtain the desired statistical power. On the other hand, the large number of patient samples required highlights the inadequate statistical power of many 'in-house' LTLV evaluation protocols.

Other Published Evaluation Protocols

CLSI EP26-A may be considered to be too burdensome for use; however some laboratories have published their own LTLV evaluation protocols which tend to be more straightforward. Katzman *et al.* published a comparison of their current LTLV protocol in use at the Mayo Clinic with CLSI EP26-A.[19] The Mayo Clinic protocol is a direct comparison between lots using 20 patient samples across the analytical range of the assay. Statistical analysis is carried out by Passing-Bablok regression, with the acceptance criteria being:

- regression line gradient between 0.9 and 1.1
- Y-intercept of regression line <50% of the lowest reportable concentration
- R^2 coefficient of determination >0.95
- mean difference between reagent lots <10%.

Both protocols were used for twelve lot change events involving six immunoassay analytes. The protocols were in agreement in nine of the evaluations, with the CLSI protocol failing lots which were passed by the in-house protocol in two cases, and the in-house protocol failing one lot which was passed by the CLSI protocol. In this case, the CLSI protocol missed a calibration bias at the upper end of the analytical range, although it does not appear that the CLSI recommendation to assess a minimum of three patient samples dispersed across the target concentrations was adhered to.

This study once again highlights a number of practical issues with CLSI EP26-A. The authors found that in some cases their imprecision data and desired critical differences prevented them from using the lookup tables in the protocol. In other cases, the lookup tables required a prohibitively large number of samples (almost three months' worth of collections in the case of one analyte).

It could also be argued that CLSI EP26-A demonstrates that the Mayo Clinic protocol was often underpowered to detect clinically meaningful change, as evidenced by the two occasions where CLSI EP26-A failed a lot which had been passed by the in-house approach. This is acknowledged by the authors who have modified their own protocol to increase its statistical power by increasing the number of samples used at medical decision limits.

The LTLV evaluation protocol in use at McMaster University has also been published.[20] The author describes his evaluation as a 'practical and useful alternative' to CLSI EP26-A. Ten patient samples are used for analysis (three samples with results at the low end of the assay range, three in the middle range, and three at the upper range, as well as one extreme result). Acceptance criteria are based on biological variation data. The preferred target is that all differences are less than one-third of the TEa. When this target is considered to be too tight, alternative targets include a comparison of the bias with the total TEa, and a review of the slope of a regression line between the two methods. In some cases, 'professional opinion' is used to determine bespoke targets.

Whilst the author of the McMaster protocol reports that it provides useful information about assay performance, he does not comment on how his method compares with CLSI EP26-A or any other method. Without data on the statistical power provided by the protocol, it is difficult to evaluate its effectiveness in picking up important variation between lots. Furthermore, like all other protocols evaluated so far, the McMaster protocol would not be expected to identify cumulative shifts in reagent performance over several lot changes.

Finally, we will conclude this section with a review of the LTLV evaluation process in current use at our own laboratory. A minimum of 5 patient samples are included in the evaluation, although ideally 20 samples are used. Samples should span the analytical range, and include extreme results. These are selected from stored samples that have been recently analysed. Results are entered into a spreadsheet template which automatically generates a Passing-Bablok regression equation and Bland-Altman plot for the entered results. Three statistical comparisons are carried out:

- a paired t-test between the two lots – Is the average relative difference significantly different from zero (alpha is set at 0.05)?
- a comparison based on biological variation – Is the difference between the two sets of results <33% of the within-subject biological variation (CV_i)?
- a comparison based upon the Royal College of Pathologists of Australasia Quality Assurance Programs (RCPAQAP) analytical performance specifications (APS)[21] – Do any of the individual differences exceed these specifications?

The new lot is considered acceptable if all three of the above questions are answered in the negative. As a statistically significant difference does not necessarily imply a clinically significant difference, we would normally accept a lot when the first criterion fails but the second two criteria pass. Any other combination of outcomes would trigger further investigation involving senior scientists and pathologists, and it may be that the lot is rejected for use.

We have found that our assessment protocol is relatively straightforward and allows for evaluation to be completed in a timely manner. The spreadsheet is linked to a database of biological variation and RCPAQAP APS information. Users simply have to select the analyte from a drop-down box and performance specifications are automatically updated. Our spreadsheet also includes a section for a separate evaluation of IQC results, which indicates whether the IQC target should be updated for the new lot of reagent. The completed document is then reviewed, signed off by a pathologist and stored electronically.

Our evaluation process does suffer from many of the same limitations as other published protocols. It has been shown that many protocols in current use are significantly underpowered[5,19,22] and this may also be true for our protocol for some analytes. Certainly, in cases where we are only able to achieve the minimum five-patient comparison, there is a substantial risk that our evaluation is underpowered as we try to balance what is statistically robust and what is practical in a busy hospital laboratory. It is also worth noting that we would not expect to detect gradual shifts in reagent or calibrator performance over several lot changes using this protocol.

Troubleshooting Lot Evaluation Failures

There are many reasons that a LTLV evaluation may fail. A failure does not immediately indicate that the new lot is unsuitable for use, although it must not be used until troubleshooting has taken place. If the failure is due to a single aberrant result it may be worth repeating the measurement (for both lots) on that particular sample. If the discrepancy is not present on repeat analysis then the new lot may be considered acceptable, although thought should be given as to likelihood of a repeat result occurring on a patient sample, and further investigation may be warranted.

Problems with calibration may also result in the failure of an evaluation, and so the calibration of each lot should be checked whenever a failure occurs. An evaluation may also fail due to the poor stability of some analytes (for example, bicarbonate), where the quantity of analyte in the sample has changed between the two measurements. The time between analysis for each lot should be kept as short as possible and samples should be capped between each measurement.

Persistent failure of one sample may indicate the presence of an interfering substance. The manufacturer's instructions for use will normally include information about previously identified analytical interferences with the assay, and the sample should be double-checked to ensure that it is appropriate for use. Even if the sample seems appropriate, occult interferences are relatively common. If possible, a second sample of a similar concentration should be substituted; an acceptable result in this sample is suggestive of an interference in the original sample. An interference may also be indicated by an abnormal appearance to the reaction profile or trace typically generated by automated instruments during sample analysis.

If all or many samples fail the evaluation, it is worth ensuring that the acceptance criteria are appropriate. It may be that the criteria represent a standard which is not achievable with current technology. Consideration may be given as to whether the criteria can be loosened without impacting on clinical outcomes. These decisions should be made by senior laboratory staff and should also involve clinicians wherever possible.

Finally, a failed evaluation may prompt laboratory staff to compare their results with those of other laboratories using the same method. This can be facilitated by the sharing of evaluation data between laboratories; assay manufacturers may also play a role in this process. A failure at one laboratory which has not been replicated elsewhere may be due to an under-performing analyser.

In some cases, a cause for clinically significant LTLV cannot be identified or addressed. In these circumstances the lot must be rejected. The laboratory should communicate this decision to the assay manufacturer, who would be expected to investigate further and may be able to provide an alternative lot for evaluation.

Limitations of Current Evaluation Procedures and Suggestions for Future Development

One of the greatest challenges when developing an evaluation protocol is achieving an acceptable statistical power. Many processes in current use do not include an evaluation of the statistical power they attain. The lookup tables in CLSI EP26-A require in excess of 20 samples to be used in some cases – a practice which is not common in most laboratories. Algeciras-Schimnich *et al.* highlight an example where an underpowered evaluation method led to an undetected shift in results for IGF-1.[5] Even when the correct number of samples has been determined, it may not be possible to achieve this number in smaller laboratories and for analytes which are infrequently requested. In this particular case, more than 100 samples would have been required to reach a statistical power of 90%.

Another problem with current protocols is that they are unable to identify smaller, cumulative shifts in reagent performance. Individually, each shift may not be clinically significant. However, if several shifts occur in the same direction, a gradual drift away from the acceptable range of results can occur, which ultimately becomes clinically significant. This may appear on EQA reports as a gradual migration of a method group away from the target value. However, if one manufacturer dominates the market then the EQA median value may also shift, masking the change in results.[23]

One approach which may identify cumulative shifts is the continuous monitoring of average results over the course of several lot changes, often referred to as the moving patient average. In the case described by Algeciras-Schimnich *et al.*, retrospective analyses of mean and median results identified a problem with performance.[5] Had such an analysis taken place prospectively, it is reasonable to assume that the issues with IGF-1 would have been identified earlier. A moving averages protocol evaluated by Van Houcke *et al.* detected several shifts in performance due to lot changes.[24] Whilst evaluation of moving averages is a useful additional approach to identifying LTLV, it may not be appropriate for analytes which display seasonal variability, such as vitamin D. This approach also assumes that the characteristics of the population being tested remain stable over time. Use of moving averages requires software support and each test must be optimised with respect to the window of samples being evaluated (batch size), filtering conditions, truncation limits, and control limits.[25] Therefore, despite the growing interest in this approach, it can be difficult to implement, especially for small laboratories with limited resources and lower throughput. Notwithstanding such limitations, measurement of the moving patient average does have the potential to strengthen our ability to detect changes in performance between lots.

Another approach to identifying cumulative shifts in performance is the comparison of regression equations obtained by successive evaluations. Such a protocol is outlined by Liu *et al.*[22] Using a statistical simulation and applying to historical data, the authors demonstrate the superiority of this approach to a 'typical' LTLV protocol, in which only the current and new lots are evaluated. In principle the approach has great appeal but application to a large number of analytes in a resource-limited environment has yet to be demonstrated.

While laboratories may take steps to manage LTLV, the preferred approach is to minimise LTLV. Therefore, an opportunity exists for the profession to collaborate with manufacturers to develop and adopt universal clinically-based LTLV acceptance criteria for individual assay reagents and calibrators. Similarly, we would urge manufacturers to implement longitudinal monitoring of trends in reagent performance, and to put in place procedures for communicating such trends to customers as soon as they become apparent. This would lighten the burden on the already stretched resources of pathology laboratories, and will significantly reduce the risk of inaccurate results being released by an unsuspecting laboratory.

Finally, LTLV evaluation can be improved by the sharing of data between laboratories. Most modern automated analysers allow the manufacturer to upload de-identified data to a central server. We would encourage manufacturers to make this data available to participating laboratories so that they may better evaluate the performance of their own assays. Alternatively, laboratories within a network may wish to create a similar centralised repository for patient results. Such an approach may be particularly helpful for tests which are infrequently requested.[23]

Conclusion

As the practice of medicine advances, clinicians increasingly expect that the results produced by the hospital laboratory are accurate and of good quality. LTLV is a potential source of inaccuracy and is best addressed by increasing collaboration between laboratories and assay manufacturers. However, until this occurs, the ongoing onus remains on the laboratory to evaluate all new lots of reagent and calibrator before putting them into use.

Each laboratory faces its own set of challenges in terms of the population served, workload and resources. This means that there can be no 'one size fits all' LTLV evaluation approach. Rather, the laboratory must develop an approach which is acceptable for its own specific needs and which provides an acceptable level of statistical robustness.

All current published LTLV evaluation processes have limitations. There is a need for a concerted effort to address these limitations, which may involve the use of moving patient averages and other strategies. Collaboration between laboratories has a key role to play if we are to successfully overcome the challenge posed by LTLV.

Acknowledgements: We would like to acknowledge and thank Mr John Calleja for his review of this manuscript.

Competing Interests: None declared.

References

1. Hölzel W. Analytical variation in immunoassays and its importance for medical decision making. Scand J Clin Lab Invest Suppl 1991;205:113-9.
2. Kim HS, Kang HJ, Whang DH, Lee SG, Park MJ, Park JY, *et al.* Analysis of reagent lot-to-lot comparability tests in five immunoassay items. Ann Clin Lab Sci 2012;42:165-73.
3. Thaler MA, Iakoubov R, Bietenbeck A, Luppa PB. Clinically relevant lot-to-lot reagent difference in a commercial immunoturbidimetric assay for glycated hemoglobin A1c. Clin Biochem 2015;48:1167-70.
4. Sandberg S, Fraser CG, Horvath AR, Jansen R, Jones G, Oosterhuis W, *et al.* Defining analytical performance specifications: Consensus Statement from the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine. Clin Chem Lab Med 2015;53:833-5.
5. Algeciras-Schimnich A, Bruns DE, Boyd JC, Bryant SC, La Fortune KA, Grebe SK. Failure of current laboratory protocols to detect lot-to-lot reagent differences: findings and possible solutions. Clin Chem 2013;59:1187-94.
6. Siemens Healthcare Diagnostics. All IMMULITE Platforms for IGF-I Shift in Patient Medians and Supply Disruption. Urgent Field Safety Notice #4005. November 2012. http://webarchive.nationalarchives.gov.uk/20150113101428/http://www.mhra.gov.uk/home/groups/fsn/documents/fieldsafetynotice/con207161.pdf (Accessed 30 August 2018).
7. Australian Commission on Safety and Quality in Health Care. 2016. Review of serious failures in reported test results for prostate-specific antigen (PSA) testing of patients by SA Pathology. https://www.hcasa.asn.au/documents/273-psa-review/file (Accessed 30 August 2018).
8. Böttcher S, van der Velden VH, Villamor N, Ritgen M, Flores-Montero J, Murua Escobar H, *et al.* Lot-to-lot stability of antibody reagents for flow cytometry. J Immunol Methods 2017;S0022-1759(17)30075-3.
9. Kitchen AD, Newham JA. Lot release testing of serological infectious disease assays used for donor and donation screening. Vox Sang 2010;98:508-16.
10. International Organization for Standardization. ISO 15189 Medical laboratories - Requirements for quality and competence. http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=56115 (Accessed 30 August 2018).
11. Palmer-Toy DE, Wang E, Winter WE, Soldin SJ, Klee GG, Howanitz JH, *et al.* Comparison of pooled fresh frozen serum to proficiency testing material in College of American Pathologists surveys: cortisol and immunoglobulin E. Arch Pathol Lab Med 2005;129:305-9.
12. Bock JL, Endres DB, Elin RJ, Wang E, Rosenzweig B, Klee GG. Comparison of fresh frozen serum to traditional proficiency testing material in a College of American Pathologists survey for ferritin, folate, and vitamin B12. Arch Pathol Lab Med 2005;129:323-7.
13. Schreiber WE, Endres DB, McDowell GA, Palomaki GE, Elin RJ, Klee GG, *et al.* Comparison of fresh frozen serum to proficiency testing material in College of American Pathologists surveys: alpha-fetoprotein, carcinoembryonic antigen, human chorionic gonadotropin, and prostate-specific antigen. Arch Pathol Lab Med 2005;129:331-7.
14. Kristensen GB, Rustad P, Berg JP, Aakre KM. Analytical Bias Exceeding Desirable Quality Goal in 4 out of 5 Common Immunoassays: Results of a Native Single Serum Sample External Quality Assessment Program for Cobalamin, Folate, Ferritin, Thyroid-Stimulating Hormone, and Free T4 Analyses. Clin Chem 2016;62:1255-63.
15. Miller WG, Erek A, Cunningham TD, Oladipo O, Scott MG, Johnson RE. Commutability limitations influence quality control results with different reagent lots. Clin Chem 2011;57:76-83.
16. Martindale RA, Cembrowski GS, Journalt LJ, Crawford JL, Tran C, Hofer TL, Rintoul BJ, *et al.* Validating New Reagents: Roadmaps Through the Wilderness. Laboratory Medicine 2006;37:347-51.
17. Kenny D, Fraser CG, Hyltoft Petersen P, Kallner A. Consensus agreement. Scand J Clin Lab Invest 1999;59:585.
18. Clinical and Laboratory Standards Institute. User Evaluation of Between-Reagent Lot Variation; Approved Guideline. CLSI document EP26-A. Wayne, PA,USA: CLSI; 2013.
19. Katzman BM, Ness KM, Algeciras-Schimnich A. Evaluation of the CLSI EP26-A protocol for detection of reagent lot-to-lot differences. Clin Biochem 2017;50:768-71.
20. Don-Wauchope AC. Lot change for reagents and

calibrators. Clin Biochem 2016;49:1211-2.
21. The Royal College of Pathologists of Australasia Quality Assurance Program. Data Analysis and Assessment Criteria Handbook. Sydney:RCPAQAP; 2018.
22. Liu J, Tan CH, Loh TP, Badrick T. Detecting long-term drift in reagent lots. Clin Chem 2015;61:1292-8.
23. Bais R, Chesher D. More on lot-to-lot changes. Clin Chem 2014;60:413-4.
24. Van Houcke SK, Stepman HC, Thienpont LM, Fiers T, Stove V, Couck P, *et al.* Long-term stability of laboratory tests and practical implications for quality management. Clin Chem Lab Med 2013;51:1227-31.
25. van Rossum HH, Kemperman H. A method for optimization and validation of moving average as continuous analytical quality control instrument demonstrated for creatinine. Clin Chim Acta 2016;457:1-7.